Chapter 61
The SURVEYMEANS Procedure

Chapter Table of Contents

OVERVIEW 3183

GETTING STARTED 3183
Simple Random Sampling 3183
Stratified Sampling 3186
Create Output Data Set 3188

SYNTAX 3189
PROC SURVEYMEANS Statement 3190
BY Statement 3193
CLASS Statement 3194
CLUSTER Statement 3195
STRATA Statement 3195
VAR Statement 3196
WEIGHT Statement 3196

DETAILS 3196
Specification of Population Totals and Sampling Rates 3196
Missing Values 3197
Primary Sampling Units (PSUs) 3198
Statistical Computation 3198
Output Data Sets 3202
Displayed Output 3203
ODS Table Names 3205

EXAMPLES 3205
Example 61.1 Stratified Cluster Sample Design 3205
Example 61.2 Unequal Weighting 3209

REFERENCES 3213

Chapter 61
The SURVEYMEANS Procedure

Overview

The SURVEYMEANS procedure produces estimates of survey population means and totals from sample survey data. The procedure also produces variance estimates, confidence limits, and other descriptive statistics. When computing these estimates, the procedure takes into account the sample design used to select the survey sample. The sample design can be a complex survey sample design with stratification, clustering, and unequal weighting.

PROC SURVEYMEANS uses the Taylor expansion method to estimate sampling errors of estimators based on complex sample designs. This method obtains a linear approximation for the estimator and then uses the variance estimate for this approximation to estimate the variance of the estimate itself (Woodruff 1971, Fuller 1975). When there are clusters, or primary sampling units (PSUs), in the sample design, the procedure estimates variance from the variation among PSU. When the design is stratified, the procedure pools stratum variance estimates to compute the overall variance estimate.

PROC SURVEYMEANS uses the Output Delivery System (ODS) to place results in output data sets. This is a departure from older SAS procedures that provide OUTPUT statements for similar functionality.

Getting Started

This section demonstrates how you can use the SURVEYMEANS procedure to produce descriptive statistics from sample survey data. For a complete description of PROC SURVEYMEANS, please refer to the “Syntax” section on page 3189. The “Examples” section on page 3205 provides more complicated examples to illustrate the applications of PROC SURVEYMEANS.

Simple Random Sampling

This example illustrates how you can use PROC SURVEYMEANS to estimate population means and proportions from sample survey data. The study population is a junior high school with a total of 4,000 students in grades 7, 8, and 9. Researchers want to know how much these students spend weekly for ice cream, on the average, and what percentage of students spend at least $10 weekly for ice cream.

To answer these questions, 40 students were selected from the entire student population using simple random sampling (SRS). Selection by simple random sampling means that all students have an equal chance or being selected, and no student can be selected more than once. Each student selected for the sample was asked how much he spends for ice cream per week, on the average. The SAS data set named IceCream saves the responses of the 40 students.

```
data IceCream;
   input Grade Spending @@;
   if (Spending < 10) then Group=’less’;
     else Group=’more’;
   datalines;
7 7  7  7  8 12  9 10  7  1  7 10  7  3  8 20  8 19  7 2
7 2  9 15  8 16  7  6  7  6  7  6  9 15  8 17  8 14  9 8
9 8  9  7  7  3  7 12  7  4  9 14  8 18  9  9  7  2  7 1
7 4  7 11  9  8  8 10  8 13  7  2  9  6  9 11  7  2  7 9
;
```

The variable Grade contains a student’s grade. The variable Spending contains a student’s response on how much he spends per week for ice cream, in dollars. The variable Group is created to indicate whether a student spends at least $10 weekly for ice cream: Group=’more’ if a student spends at least $10, or Group=’less’ if a student spends less than $10.

You can use PROC SURVEYMEANS to produce estimates for the entire student population, based on this random sample of 40 students.

```
title1 ’Analysis of Ice Cream Spending’;
title2 ’Simple Random Sampling Design’;
proc surveymeans data=IceCream total=4000;
   var Spending Group;
   run;
```

The PROC SURVEYMEANS statement invokes the procedure. The TOTAL=4000 option specifies the total number of students in the study population, or school. The procedure uses this total to adjust variance estimates for the effects of sampling from a finite population. The VAR statement names the variables to analyze, Spending and Group.

```
                          Analysis of Ice Cream Spending
                          Simple Random Sampling Design

                            The SURVEYMEANS Procedure

                                   Data Summary

                       Number of Observations            40


                             Class Level Information

              Class
              Variable      Levels    Values

              Group              2    less more


                                    Statistics

                                               Std Error     Lower 95%     Upper 95%
Variable                  N          Mean        of Mean   CL for Mean   CL for Mean
------------------------------------------------------------------------------------
Spending                 40      8.750000       0.845139      7.040545     10.459455
Group = less             23      0.575000       0.078761      0.415690      0.734310
Group = more             17      0.425000       0.078761      0.265690      0.584310
------------------------------------------------------------------------------------
```

Figure 61.1. Analysis of Ice Cream Spending, Simple Random Sampling Design

Figure 61.1 displays the results from this analysis. There are a total of 40 observations used in the analysis. The "Class Level Information" table lists the two levels of the variable Group. This variable is a character variable, and so PROC SURVEYMEANS provides a categorical analysis for it, estimating the relative frequency or proportion for each level. If you want a categorical analysis for a numeric variable, you can name that variable in the CLASS statement.

The "Statistics" table displays the estimates for each analysis variable. By default, PROC SURVEYMEANS displays the number of observations, the estimate of the mean, its standard error, and 95% confidence limits for the mean. You can obtain other statistics by specifying the corresponding statistic-keywords in the PROC SURVEYMEANS statement.

The estimate of the average weekly ice cream expense is $8.75 for students in this school. The standard error of this estimate if $0.84, and the 95% confidence interval for weekly ice cream expense is from $7.04 to $10.46.

The analysis variable Group is a character variable, and so PROC SURVEYMEANS analyzes it as categorical, estimating the relative frequency or proportion for each level or category. These estimates are displayed in the Mean column of the "Statistics" table. It is estimated that 57.5% of all students spend less than $10 weekly on ice cream, while 42.5% of the students spend at least $10 weekly. The standard error of each estimate is 7.9%.

Stratified Sampling

Suppose that the sample of students described in the previous section was actually selected using stratified random sampling. In stratified sampling, the study population is divided into nonoverlapping strata, and samples are selected independently from each stratum.

The list of students in this junior high school was stratified by grade, yielding three strata: grades 7, 8, and 9. A simple random sample of students was selected from each grade. Table 61.1 shows the total number of students in each grade.

Table 61.1. Number of Students by Grade

Grade	Number of Students
7	1,824
8	1,025
9	1,151
Total	4,000

To analyze this stratified sample from a finite population, you need to provide the population totals for each stratum to PROC SURVEYMEANS. The SAS data set named StudentTotal contains the information from Table 61.1.

```
data StudentTotal;
   input Grade _total_; datalines;
7 1824
8 1025
9 1151
;
```

The variable Grade is the stratum identification variable, and the variable _TOTAL_ contains the total number of students for each stratum. PROC SURVEYMEANS requires you to store the stratum population totals in a variable named _TOTAL_.

The procedure uses the stratum population totals to adjust variance estimates for the effects of sampling from a finite population. If you do not provide population totals or sampling rates, then the procedure assumes that the proportion of the population in the sample is very small, and it does not include a finite population correction in the computations.

The following SAS statements perform the analysis of the survey data.

```
title1 'Analysis of Ice Cream Spending';
title2 'Stratified Simple Random Sampling Design';
proc surveymeans data=IceCream total=StudentTotal;
   stratum Grade / list;
   var Spending Group;
run;
```

The PROC SURVEYMEANS statement invokes the procedure. The DATA= option names the SAS data set IceCream as the input data set to be analyzed. The TOTAL= option names the data set StudentTotal as the input data set containing the stratum population totals. Comparing this to the analysis in the “Simple Random Sampling” section on page 3183, notice that the TOTAL=StudentTotal option is used here instead of the TOTAL=4000 option. In this stratified sample design, the population totals are different for different strata, and so they are provided to PROC SURVEYMEANS in a SAS data set.

The STRATA statement identifies the stratification variable Grade. The LIST option in the STRATA statement requests that the procedure display stratum information.

```
                   Analysis of Ice Cream Spending
               Stratified Simple Random Sampling Design

                      The SURVEYMEANS Procedure

                             Data Summary

                 Number of Strata                   3
                 Number of Observations            40


                       Class Level Information

       Class
       Variable      Levels    Values

       Group              2    less more
```

Figure 61.2. Data Summary

Figure 61.2 provides information on the input data set. There are three strata in the design, and 40 observations in the sample. The categorical variable Group has two levels, ’less’ and ’more’.

```
                   Analysis of Ice Cream Spending
               Stratified Simple Random Sampling Design

                      The SURVEYMEANS Procedure

                         Stratum Information

Stratum              Population   Sampling
 Index       Grade        Total       Rate      N Obs   Variable              N
-------------------------------------------------------------------------------
   1             7         1824       0.01         20   Spending             20
                                                        Group = less         17
                                                        Group = more          3
   2             8         1025       0.01          9   Spending              9
                                                        Group = less          0
                                                        Group = more          9
   3             9         1151       0.01         11   Spending             11
                                                        Group = less          6
                                                        Group = more          5
-------------------------------------------------------------------------------
```

Figure 61.3. Stratum Information

Figure 61.3 displays information for each stratum. The table displays a Stratum Index and the values of the STRATA variable. The Stratum Index identifies each stratum by a sequentially-assigned number. For each stratum, the table gives the population total (total number of students), the sampling rate, and the sample size. The stratum sampling rate is the ratio of the number of students in the sample to the number of students in the population for that stratum. The table also lists each analysis variable and the number of stratum observations for that variable. For categorical variables, the table lists each level and the number of sample observations in that level.

```
                          Analysis of Ice Cream Spending
                     Stratified Simple Random Sampling Design

                            The SURVEYMEANS Procedure

                                    Statistics

                                             Std Error     Lower 95%     Upper 95%
Variable                 N          Mean       of Mean   CL for Mean   CL for Mean
----------------------------------------------------------------------------------
Spending                40      8.750000      0.530531      7.675043      9.824957
Group = less            23      0.575000      0.059299      0.454850      0.695150
Group = more            17      0.425000      0.059299      0.304850      0.545150
----------------------------------------------------------------------------------
```

Figure 61.4. Ice Cream Spending Analysis, Stratified SRS Design

Figure 61.4 shows that the estimate of average weekly ice cream expense is $8.75 for students in this school, with a standard error of $0.53, and a 95% confidence interval from $7.68 to $9.82. The mean estimate of $8.75 is the same as the one shown in Figure 61.1, which was computed under the assumption of a simple random sampling design without stratification. However, the standard error computed for the stratified design is $0.53, which is less than the standard error of $0.84 shown in Figure 61.1.

Figure 61.4 shows that an estimate of 57.5% of all students spend less than $10 weekly on ice cream, and 42.5% spend more, with a standard error of 0.06%.

Create Output Data Set

PROC SURVEYMEANS uses the Output Delivery System (ODS) to create output data sets. This is a departure from older SAS procedures that provide OUTPUT statements for similar functionality. For more information on ODS, see Chapter 15, “Using the Output Delivery System.”

For example, to place the “Statistics” table shown in Figure 61.4 in the previous section in an output data set, you use the ODS OUTPUT statement as follows:

```
title1 ’Analysis of Ice Cream Spending’;
title2 ’Stratified Simple Random Sampling Design’;
proc surveymeans data=IceCream total=StudentTotal;
   stratum Grade / list;
   var Spending Group;
   ods output Statistics=MyStat;
run;
```

The statement

```
ods output Statistics=MyStat;
```

requests that the "Statistics" table that appears in Figure 61.4 be placed in a SAS data set named MyStat.

The PRINT procedure displays observations of the data set MyStat:

```
proc print data=MyStat;
run;
```

Figure 61.4 displays the observations in the data set MyStat.

```
                         Analysis of Ice Cream Spending
                     Stratified Simple Random Sampling Design


    Variable_       N_For_     Mean_For_  StdErr_For_ LowerCLMean_ UpperCLMean_
Obs Spending      Spending      Spending     Spending For_Spending For_Spending

 1  Spending            40      8.750000     0.530531     7.675043     9.824957

                                  Mean_For_  StdErr_For_ LowerCLMean_ UpperCLMean_
      Level_1_   N_For_Level_      Level_1_     Level_1_  For_Level_   For_Level_
Obs   Of_Group     1_Of_Group      Of_Group     Of_Group  1_Of_Group   1_Of_Group

 1  Group = less           23      0.575000     0.059299     0.454850     0.695150

                                  Mean_For_  StdErr_For_ LowerCLMean_ UpperCLMean_
      Level_2_   N_For_Level_      Level_2_     Level_2_  For_Level_   For_Level_
Obs   Of_Group     2_Of_Group      Of_Group     Of_Group  2_Of_Group   2_Of_Group

 1  Group = more           17      0.425000     0.059299     0.304850     0.545150
```

Figure 61.5. The Data Set MyStat

Syntax

The following statements are available in PROC SURVEYMEANS.

> **PROC SURVEYMEANS** < *options* > < *statistic-keywords* > **;**
>   **BY** *variables* **;**
>   **CLASS** *variables* **;**
>   **CLUSTER** *variables* **;**
>   **STRATA** *variables* < / *option* > **;**
>   **VAR** *variables* **;**
>   **WEIGHT** *variable* **;**

The PROC SURVEYMEANS statement invokes the procedure. It optionally names the input data sets and specifies statistics for the procedure to compute. The PROC SURVEYMEANS statement is required.

The VAR statement identifies the variables to be analyzed. The CLASS statement identifies those numeric variables that are to be analyzed as categorical variables. The STRATA statement lists the variables that form the strata in a stratified sample design. The CLUSTER statement specifies cluster identification variables in a clustered sample design. The WEIGHT statement names the sampling weight variable. You can use a BY statement with PROC SURVEYMEANS to obtain separate analyses for groups defined by the BY variables.

All statements can appear multiple times except the PROC SURVEYMEANS statement and the WEIGHT statement, which can appear only once.

The rest of this section gives detailed syntax information for the BY, CLASS, CLUSTER, STRATA, VAR, and WEIGHT statements in alphabetical order after the description of the PROC SURVEYMEANS statement.

PROC SURVEYMEANS Statement

PROC SURVEYMEANS < *options* > < *statistic-keywords* > **;**

The PROC SURVEYMEANS statement invokes the procedure. In this statement, you identify the data set to be analyzed and specify sample design information. The DATA= option names the input data set to be analyzed. If your analysis includes a finite population correction factor, you can input either the sampling rate or the population total using the RATE= or TOTAL= option. If your design is stratified, with different sampling rates or totals for different strata, then you can input these stratum rates or totals in a SAS data set containing the stratification variables.

In the PROC SURVEYMEANS statement, you also can use *statistic-keywords* to specify statistics for the procedure to compute. Available statistics include the population mean and population total, together with their variance estimates and confidence limits. You can also request data set summary information and sample design information.

You can specify the following options in the PROC SURVEYMEANS statement.

ALPHA=α

sets the confidence level for confidence limits. The value of the ALPHA= option must be between 0.0001 and 0.9999, and the default value is 0.05. A confidence level of α produces $100(1 - \alpha)\%$ confidence limits. The default of ALPHA=0.05 produces 95% confidence limits. If α is between 0 and 1 but outside the range of 0.0001 to 0.9999, the procedure uses the closest range endpoint. For example, if you specify ALPHA=0.000001, the procedure uses 0.0001 to determine confidence limits.

DATA=*SAS-data-set*

specifies the SAS data set to be analyzed by PROC SURVEYMEANS. If you omit the DATA= option, the procedure uses the most recently created SAS data set.

MISSING

requests that the procedure treat missing values as a valid category for categorical variables.

ORDER=DATA | FORMATTED | INTERNAL

specifies the order in which the values of the categorical variables are to be reported. Note that the ORDER= option applies to all the categorical variables. The exception is ORDER=FORMATTED (the default) for numeric variables for which you have supplied no explicit format (that is, for which there is no corresponding FORMAT statement in the current PROC SURVEYMEANS run or in the DATA step that created the data set). In this case, the values of the numerical categorical variables are ordered by their internal (numeric) value. The following shows how PROC SURVEYMEANS interprets values of the ORDER= option.

DATA — orders values according to their order in the input data set.

FORMATTED — orders values by their formatted values. This order is operating environment dependent. By default, the order is ascending.

INTERNAL — orders values by their unformatted values, which yields the same order that the SORT procedure does. This order is operating environment dependent.

By default, ORDER=FORMATTED.

RATE=*value* | *SAS-data-set*
R=*value* | *SAS-data-set*

specifies the sampling rate as a positive *value*, or names an input data set that contains the stratum sampling rates. The procedure uses this information to compute a finite population correction for variance estimation. If your sample design has multiple stages, you should specify the *first-stage sampling rate*, which is the ratio of the number of PSUs selected to the total number of PSUs in the population

For a nonstratified sample design, or for a stratified sample design with the same sampling rate in all strata, you should specify a positive *value* for the RATE= option. If your design is stratified with different sampling rates in the strata, then you should name a SAS data set that contains the stratification variables and the sampling rates. See the section "Specification of Population Totals and Sampling Rates" on page 3196 for details.

The sampling rate *value* must be a positive number. You can specify *value* as a number between 0 and 1. Or you can specify *value* in percentage form as a number between 1 and 100, and PROC SURVEYMEANS will convert that number to a proportion. The procedure treats the value 1 as 100%, and not the percentage form 1%.

If you do not specify the TOTAL= option or the RATE= option, then the variance estimation does not include a finite population correction. You cannot specify both the TOTAL= option and the RATE= option.

TOTAL=*value* | *SAS-data-set*
N=*value* | *SAS-data-set*

specifies the total number of primary sampling units (PSUs) in the study population as a positive *value*, or names an input data set that contains the stratum population totals. The procedure uses this information to compute a finite population correction for variance estimation.

For a nonstratified sample design, or for a stratified sample design with the same population total in all strata, you should specify a positive *value* for the TOTAL= option. If your sample design is stratified with different population totals in the strata, then you should name a SAS data set that contains the stratification variables and the population totals. See the section "Specification of Population Totals and Sampling Rates" on page 3196 for details.

If you do not specify the TOTAL= option or the RATE= option, then the variance estimation does not include a finite population correction. You cannot specify both the TOTAL= option and the RATE= option.

statistic-keywords

specifies the statistics for the procedure to compute. If you do not specify any statistic-keywords, PROC SURVEYMEANS computes the NOBS, MEAN, STDERR, and CLM statistics by default.

PROC SURVEYMEANS performs univariate analysis, analyzing each variable separately. Thus the number of nonmissing and missing observations may not be the same for all analysis variables. See the section "Missing Values" on page 3197 for more information.

The statistics produced depend on the type of the analysis variable. If you name a numeric variable in the CLASS statement, then the procedure analyzes that variable as a categorical variable. The procedure always analyzes character variables as categorical. See the section "CLASS Statement" on page 3194 for more information.

PROC SURVEYMEANS computes MIN, MAX, and RANGE for numeric variables but not for categorical variables. For numeric variables, the keyword MEAN produces the mean, but for categorical variables it produces the proportion in each category or level. Also for categorical variables, the keyword NOBS produces the number of observations for each variable level, and the keyword NMISS produces the number of missing observations for each level. If you request the keyword NCLUSTER for a categorical variable, PROC SURVEYMEANS displays for each level the number of clusters with observations in that level. PROC SURVEYMEANS computes SUMWGT the same for categorical and numeric variables, as the sum of the weights over all nonmissing observations.

The valid statistic-keywords are as follows:

ALL	all statistics listed
CLM	$100(1-\alpha)\%$ confidence limits for the MEAN, where α is determined by the ALPHA= option described on page 3190, and the default is $\alpha = 0.05$

CLSUM	$100(1-\alpha)$% confidence limits for the SUM, where α is determined by the ALPHA= option described on page 3190, and the default is $\alpha = 0.05$
CV	coefficient of variation
DF	degrees of freedom for the *t* test
MAX	maximum value
MEAN	mean for a numeric variable, or the proportion in each category for a categorical variable
MIN	minimum value
NCLUSTER	number of clusters
NMISS	number of missing observations
NOBS	number of nonmissing observations
RANGE	range, MAX$-$MIN
STD	standard deviation of the SUM. When you request SUM, the procedure computes STD by default.
STDERR	standard error of the MEAN. When you request MEAN, the procedure computes STDERR by default.
SUM	weighted sum, $\sum w_i y_i$, or estimated population total when the appropriate sampling weights are used
SUMWGT	sum of the weights, $\sum w_i$
T	*t* value for H_0 : population MEAN $= 0$, and its two tailed *p*-value with DF degrees of freedom
VAR	variance of the MEAN
VARSUM	variance of the SUM

BY Statement

BY *variables* **;**

You can specify a BY statement with PROC SURVEYMEANS to obtain separate analyses on observations in groups defined by the BY variables.

Note that using a BY statement provides completely separate analyses of the BY groups. It does not provide a statistically valid subpopulation or domain analysis, where the total number of units in the subpopulation is not known with certainty. For more information on estimating the variance of subpopulation means and totals from sample survey data, refer to Cochran (1977). However, note that you can produce a domain analysis with PROC SURVEYREG (see Example 62.7 on page 3269).

When a BY statement appears, the procedure expects the input data sets to be sorted in order of the BY variables. The *variables* are one or more variables in the input data set.

If you specify more than one BY statement, the procedure uses only the latest BY statement and ignores any previous ones.

If your input data set is not sorted in ascending order, use one of the following alternatives:

- Sort the data using the SORT procedure with a similar BY statement.
- Use the BY statement options NOTSORTED or DESCENDING in the BY statement. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.
- Create an index on the BY variables using the DATASETS procedure.

For more information on the BY statement, refer to the discussion in *SAS Language Reference: Concepts*. For more information on the DATASETS procedure, refer to the discussion in the *SAS Procedures Guide*.

CLASS Statement

CLASS | **CLASSES** *variables* **;**

The CLASS statement names variables to be analyzed as categorical variables. For categorical variables, PROC SURVEYMEANS estimates the proportion in each category or level, instead of the overall mean. PROC SURVEYMEANS always analyzes character variables as categorical. If you want categorical analysis for a numeric variable, you must include that variable in the CLASS statement.

The CLASS *variables* are one or more variables in the DATA= input data set. These variables can be either character or numeric. The formatted values of the CLASS variables determine the categorical variable levels. Thus, you can use formats to group values into levels. Refer to the discussion of the FORMAT procedure in the *SAS Procedures Guide* and to the discussions of the FORMAT statement and SAS formats in *SAS Language Reference: Dictionary*.

You can use multiple CLASS statements to specify categorical variables.

When you specify class variables, the procedure uses the SAS system option SUMSIZE= for the amount of memory that is available for data analysis. Refer to the chapter on SAS System options in *SAS Language Reference: Dictionary* for a description of the SUMSIZE= option.

CLUSTER Statement

CLUSTER | **CLUSTERS** *variables* **;**

The CLUSTER statement names variables that identify the clusters in a clustered sample design. The combinations of categories of CLUSTER variables define the clusters in the sample. If there is a STRATA statement, clusters are nested within strata.

If your sample design has clustering at multiple stages, you should identify only the first-stage clusters, or primary sampling units (PSUs), in the CLUSTER statement. See the section “Primary Sampling Units (PSUs)” on page 3198 for more information.

The CLUSTER *variables* are one or more variables in the DATA= input data set described on page 3190. These variables can be either character or numeric. The formatted values of the CLUSTER variables determine the CLUSTER variable levels. Thus, you can use formats to group values into levels. Refer to the discussion of the FORMAT procedure in the *SAS Procedures Guide* and to the discussions of the FORMAT statement and SAS formats in *SAS Language Reference: Dictionary*.

You can use multiple CLUSTER statements to specify cluster variables. The procedure uses variables from all CLUSTER statements to create clusters.

STRATA Statement

STRATA | **STRATUM** *variables* < */ option* > **;**

The STRATA statement names variables that form the strata in a stratified sample design. The combinations of categories of STRATA variables define the strata in the sample.

If your sample design has stratification at multiple stages, you should identify only the first-stage strata in the STRATA statement. See the section “Specification of Population Totals and Sampling Rates” on page 3196 for more information.

The STRATA *variables* are one or more variables in the DATA= input data set described on page 3190. These variables can be either character or numeric. The formatted values of the STRATA variables determine the levels. Thus, you can use formats to group values into levels. See the discussion of the FORMAT procedure in the *SAS Procedures Guide*.

You can specify the following option in the STRATA statement after a slash (/).

LIST

displays a “Stratum Information” table, which includes values of the STRATA variables and sampling rates for each stratum. This table also provides the number of observations and number of clusters for each stratum and analysis variable. See the section “Displayed Output” on page 3203 for more details.

VAR Statement

VAR *variables* **;**

The VAR statement names the variables to be analyzed.

If you want a categorical analysis for a numeric variable, you must also name that variable in the CLASS statement. For categorical variables, PROC SURVEYMEANS estimates the proportion in each category or level, instead of the overall mean. Character variables are always analyzed as categorical variables. See the section "CLASS Statement" on page 3194 for more information.

If you do not specify a VAR statement, then PROC SURVEYMEANS analyzes all variables in the DATA= input data set, except those named in the BY, CLUSTER, STRATA, and WEIGHT statements.

WEIGHT Statement

WEIGHT | **WGT** *variable* **;**

The WEIGHT statement names the variable that contains the sampling weights. This variable must be numeric. If you do not specify a WEIGHT statement, PROC SURVEYMEANS assigns all observations a weight of 1. Sampling weights must be positive numbers. If an observation has a weight that is nonpositive or missing, then the procedure omits that observation from the analysis. If you specify more than one WEIGHT statement, the procedure uses only the first WEIGHT statement and ignores the rest.

Details

Specification of Population Totals and Sampling Rates

If your analysis should include a finite population correction (*fpc*), you can input either the sampling rate or the population total using the RATE= option or the TOTAL= option. (You cannot specify both of these options in the same PROC SURVEYMEANS statement.) If you do not specify one of these options, the procedure does not use the *fpc* when computing variance estimates. For fairly small sampling fractions, it is appropriate to ignore this correction. Refer to Cochran (1977) and Kish (1965).

If your design has multiple stages of selection and you are specifying the RATE= option, you should input the first-stage sampling rate, which is the ratio of the number of PSUs in the sample to the total number of PSUs in the study population. If you are specifying the TOTAL= option for a multistage design, you should input the total number of PSUs in the study population. See the section “Primary Sampling Units (PSUs)” on page 3198 for more details.

For a nonstratified sample design, or for a stratified sample design with the same sampling rate or the same population total in all strata, you should use the RATE=*value* option or the TOTAL=*value* option. If your sample design is stratified with different sampling rates or population totals in the strata, then you can use the RATE=*SAS-data-set* option or the TOTAL=*SAS-data-set* option to name a SAS data set that contains the stratum sampling rates or totals. This data set is called a *secondary data set*, as opposed to the *primary data set* that you specify with the DATA= option.

The secondary data set must contain all the stratification variables listed in the STRATA statement and all the variables in the BY statement. If there are formats associated with the STRATA variables and the BY variables, then the formats must be consistent in the primary and the secondary data sets. If you specify the TOTAL=*SAS-data-set* option, the secondary data set must have a variable named _TOTAL_ that contains the stratum population totals. Or if you specify the RATE=*SAS-data-set* option, the secondary data set must have a variable named _RATE_ that contains the stratum sampling rates. The secondary data set must contain all BY and STRATA groups that occur in the primary data set. If the secondary data set contains more than one observation for any one stratum, then the procedure uses the first value of _TOTAL_ or _RATE_ for that stratum and ignores the rest.

The *value* in the RATE= option or the values of _RATE_ in the secondary data set must be positive numbers. You can specify *value* as a number between 0 and 1. Or you can specify *value* in percentage form as a number between 1 and 100, and PROC SURVEYMEANS will convert that number to a proportion. The procedure treats the value 1 as 100%, and not the percentage form 1%.

If you specify the TOTAL=*value* option, *value* must not be less than the sample size. If you provide stratum population totals in a secondary data set, these values must not be less than the corresponding stratum sample sizes.

Missing Values

If an observation has a missing value or a nonpositive value for the WEIGHT variable, then PROC SURVEYMEANS excludes that observation from the analysis. An observation is also excluded if it has a missing value for any STRATA or CLUSTER variable, unless the MISSING option described on page 3191, is used.

When computing statistics for an analysis variable, PROC SURVEYMEANS omits observations with missing values for that variable. The procedure bases statistics for each variable only on observations that have nonmissing values for that variable. If you specify the MISSING option in the PROC SURVEYMEANS statement, the procedure treats missing values of a categorical variable as a valid category.

The procedure performs univariate analysis and analyzes each VAR variable separately. Thus, the number of missing observations may be different for different variables. You can specify the keyword NMISS in the PROC SURVEYMEANS statement to display the number of missing values for each analysis variable in the "Statistics" table.

If you have missing values in your survey data for any reason (such as nonresponse), this can compromise the quality of your survey results. If the respondents are different from the nonrespondents with regard to a survey effect or outcome, then survey estimates will be biased and will not accurately represent the survey population. There are a variety of techniques in sample design and survey operations that can reduce nonresponse. Once data collection is complete, you can use imputation to replace missing values with acceptable values, and you can use sampling weight adjustments to compensate for nonresponse. You should complete this data preparation and adjustment before you analyze your data with PROC SURVEYMEANS. Refer to Cochran (1977) for more details.

Primary Sampling Units (PSUs)

When you have clusters, or primary sampling units (PSUs), in your sample design, the procedure estimates variance from the variation among PSUs. See the section "Variance and Standard Error of the Mean" on page 3200 and the section "Variance and Standard Deviation of the Total" on page 3202. You can use the CLUSTER statement to identify the first stage clusters in your design. PROC SURVEYMEANS assumes that each cluster represents a PSU in the sample and that each observation is an element of a PSU. If you do not specify a CLUSTER statement, the procedure treats each observation as a PSU.

Statistical Computation

The SURVEYMEANS procedure uses the Taylor expansion method to estimate sampling errors of estimators based on complex sample designs. This method obtains a linear approximation for the estimator and then uses the variance estimate for this approximation to estimate the variance of the estimate itself (Woodruff 1971, Fuller 1975). When there are clusters, or PSUs, in the sample design, the procedure estimates variance from the variation among PSUs. When the design is stratified, the procedure pools stratum variance estimates to compute the overall variance estimate. For *t* tests of the estimates, the degrees of freedom equals the number of clusters minus the number of strata in the sample design.

For a multistage sample design, the variance estimation method depends only on the first stage of the sample design. So, the required input includes only first-stage cluster (PSU) and first-stage stratum identification. You do not need to input design information about any additional stages of sampling. This variance estimation method assumes that the first-stage sampling fraction is small, or the first-stage sample is drawn with replacement, as it often is in practice.

For more information on the analysis of sample survey data, refer to Lee, Forthoffer, and Lorimor (1989), Cochran (1977), Kish (1965), and Hansen, Hurwitz, and Madow (1953).

Definitions and Notation

For a stratified clustered sample design, together with the sampling weights, the sample can be represented by an $n \times (P+1)$ matrix

$$\begin{aligned}(\mathbf{w}, \mathbf{Y}) &= (w_{hij}, \mathbf{y}_{hij}) \\ &= \left(w_{hij}, y_{hij}^{(1)}, y_{hij}^{(2)}, \ldots, y_{hij}^{(P)}\right)\end{aligned}$$

where

- $h = 1, 2, \ldots, H$ is the stratum number, with a total of H strata
- $i = 1, 2, \ldots, n_h$ is the cluster number within stratum h, with a total of n_h clusters
- $j = 1, 2, \ldots, m_{hi}$ is the unit number within cluster i of stratum h, with a total of m_{hi} units
- $p = 1, 2, \ldots, P$ is the analysis variable number, with a total of P variables
- $n = \sum_{h=1}^{H} \sum_{i=1}^{n_h} m_{hi}$ is the total number of observations in the sample
- w_{hij} denotes the sampling weight for observation j in cluster i of stratum h
- $\mathbf{y}_{hij} = \left(y_{hij}^{(1}), y_{hij}^{(2)}, \ldots, y_{hij}^{(P)}\right)$ are the observed values of the analysis variables for observation j in cluster i of stratum h, including both the values of numerical variables and the values of indicator variables for levels of categorical variables.

For a categorical variable C, let l denote the number of levels of C, and denote the level values as $c_1, c_2, \ldots, c_l$. Then there are l indicator variables associated with these levels. That is, for level $C = c_k$ $(k = 1, 2, \ldots, l)$, a $y^{(q)}$ $(q \in \{1, 2, \ldots, P\})$ contains the values of the indicator variable for the category $C = c_k$, with the value of observation j in cluster i of stratum h:

$$y_{hij}^{(q)} = I_{\{C=c_k\}}(h,i,j) = \begin{cases} 1 & \text{if } C_{hij} = c_k \\ 0 & \text{otherwise} \end{cases}$$

Therefore, the total number of analysis variables, P, is the total number of numerical variables plus the total number of levels of all categorical variables.

Also, f_h denotes the sampling rate for stratum h. You can use the TOTAL= option or the RATE= option to input population totals or sampling rates. See the section "Specification of Population Totals and Sampling Rates" on page 3196 for details. If you input stratum totals, PROC SURVEYMEANS computes f_h as the ratio of the stratum sample size to the stratum total. If you input stratum sampling rates, PROC SURVEYMEANS uses these values directly for f_h. If you do not specify the TOTAL= option or the RATE= option, then the procedure assumes that the stratum sampling rates f_h are negligible, and a finite population correction is not used when computing variances.

This notation is also applicable to other sample designs. For example, for a sample design without stratification, you can let $H = 1$; for a sample design without clusters, you can let $m_{hi} = 1$ for every h and i.

Mean

When you specify the keyword MEAN, the procedure computes the estimate of the mean (mean per element) from the survey data. Also, the procedure computes the mean by default if you do not specify any statistic-keywords in the PROC SURVEYMEANS statement.

PROC SURVEYMEANS computes the estimate of the mean as

$$\widehat{\bar{Y}} \quad = \quad \left(\sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} \, y_{hij} \right) / \; w_{\cdots}$$

where

$$w_{\cdots} \quad = \quad \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}$$

is the sum of the weights over all observations in the sample.

Variance and Standard Error of the Mean

When you specify the keyword STDERR, the procedure computes the standard error of the mean. Also, the procedure computes the standard error by default if you specify the keyword MEAN, or if you do not specify any statistic-keywords in the PROC SURVEYMEANS statement. The keyword VAR requests the variance of the mean.

PROC SURVEYMEANS uses the Taylor series expansion theory to estimate the variance of the mean $\widehat{\bar{Y}}$. The procedure computes the estimated variance as

$$\widehat{V}(\widehat{\bar{Y}}) \quad = \quad \sum_{h=1}^{H} \frac{n_h(1-f_h)}{n_h - 1} \sum_{i=1}^{n_h} (e_{hi\cdot} - \bar{e}_{h\cdot\cdot})^2$$

where

$$e_{hi\cdot} \quad = \quad \left(\sum_{j=1}^{m_{hi}} w_{hij} \, (y_{hij} - \widehat{\bar{Y}}) \right) / \; w_{\cdots}$$

$$\bar{e}_{h\cdot\cdot} \quad = \quad \left(\sum_{i=1}^{n_h} e_{hi\cdot} \right) / \; n_h$$

The standard error of the mean is the square root of the estimated variance.

$$\text{StdErr}(\widehat{\bar{Y}}) = \sqrt{\widehat{V}(\widehat{\bar{Y}})}$$

t Test for the Mean

If you specify the keyword T, PROC SURVEYMEANS computes the *t* value for testing that the population mean equals zero, $H_0 : \bar{Y} = 0$. The test statistic equals

$$t(\widehat{\bar{Y}}) \quad = \quad \widehat{\bar{Y}} \ / \ \text{StdErr}(\widehat{\bar{Y}})$$

The two-sided *p*-value for this test is

$$\text{Prob}(\ |T| > |t(\widehat{\bar{Y}})|\ )$$

where T is a random variable with the *t* distribution with *df* degrees of freedom.

PROC SURVEYMEANS calculates the degrees of freedom for the *t* test as the number of clusters minus the number of strata. If there are no clusters, then *df* equals the number of observations minus the number of strata. If the design is not stratified, then *df* equals the number of clusters minus one. The procedure displays *df* for the *t* test if you specify the keyword DF in the PROC SURVEYMEANS statement.

Confidence Limits for the Mean

If you specify the keyword CLM, the procedure computes confidence limits for the mean. Also, the procedure includes the confidence limits by default if you do not specify any statistic-keywords in the PROC SURVEYMEANS statement.

The confidence coefficient is determined by the value of the ALPHA= option, which by default equals 0.05 and produces 95% confidence limits. The confidence limits are computed as

$$\widehat{\bar{Y}} \quad \pm \quad \text{StdErr}(\widehat{\bar{Y}}) \ \cdot \ t_{df,\ \alpha/2}$$

where $\widehat{\bar{Y}}$ is the estimate of the mean, StdErr($\widehat{\bar{Y}}$) is the standard error of the mean, and $t_{df,\ \alpha/2}$ is the $100(1 - \alpha/2)$ percentile of the *t* distribution with *df* calculated as described in the section “t Test for the Mean”.

Coefficient of Variation

If you specify the keyword CV, PROC SURVEYMEANS computes the coefficient of variation, which is the ratio of the standard error of the mean to the estimated mean.

$$cv \quad = \quad \text{StdErr}(\widehat{\bar{Y}}) \ / \ \widehat{\bar{Y}}$$

Proportions

If you specify the keyword MEAN for a categorical variable, PROC SURVEYMEANS estimates the proportion, or relative frequency, for each level of the categorical variable. If you do not specify any statistic-keywords in the PROC SURVEYMEANS statement, the procedure estimates the proportions for levels of the categorical variables, together with their standard errors and confidence limits.

The procedure estimates the proportion in level c_k for variable C as

$$\hat{p} = \frac{\sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \ w_{hij} \ y_{hij}^{(q)}}{\sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \ w_{hij}}$$

where $y_{hij}^{(q)}$ is value of the indicator function for level $C = c_k$, defined in the section "Definitions and Notation" on page 3199, $y_{hij}^{(q)}$ equals 1 if the observed value of variable C equals c_k, and $y_{hij}^{(q)}$ equals 0 otherwise. Since the proportion estimator is actually an estimator of the mean for an indicator variable, the procedure computes its variance and standard error according to the method outlined in the section "Variance and Standard Error of the Mean" on page 3200. Similarly, the procedure computes confidence limits for proportions as described in the section "Confidence Limits for the Mean" on page 3201.

Total

If you specify the keyword SUM, the procedure computes the estimate of the population total from the survey data. The estimate of the total is the weighted sum over the sample.

$$\widehat{Y} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}\, y_{hij}$$

For a categorical variable level, $\widehat{Y}$ estimates its total frequency in the population.

Variance and Standard Deviation of the Total

When you specify the keyword STD or the keyword SUM, the procedure estimates the standard deviation of the total. The keyword VARSUM requests the variance of the total.

PROC SURVEYMEANS estimates the variance of the total as

$$\widehat{V}(\widehat{Y}) = \sum_{h=1}^{H} \frac{n_h(1-f_h)}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi\cdot} - \bar{y}_{h\cdot\cdot})^2$$

where

$$\begin{aligned} y_{hi\cdot} &= \sum_{j=1}^{m_{hi}} w_{hij}\, y_{hij} \\ \bar{y}_{h\cdot\cdot} &= \left(\sum_{i=1}^{n_h} y_{hi\cdot} \right) / \, n_h \end{aligned}$$

The standard deviation of the total equals

$$\text{Std}(\widehat{Y}) = \sqrt{\widehat{V}(\widehat{Y})}$$

Confidence Limits of a Total

If you specify the keyword CLSUM, the procedure computes confidence limits for the total. The confidence coefficient is determined by the value of the ALPHA= option, which by default equals 0.05 and produces 95% confidence limits. The confidence limits are computed as

$$\widehat{Y} \quad \pm \quad \text{Std}(\widehat{Y}) \ \cdot \ t_{df,\, \alpha/2}$$

where $\widehat{Y}$ is the estimate of the total, $\mathrm{Std}(\widehat{Y})$ is the estimated standard deviation, and $t_{df,\ \alpha/2}$ is the $100(1-\alpha/2)$ percentile of the *t* distribution with *df* calculated as described in the section "t Test for the Mean" on page 3200.

Output Data Sets

Output data sets from PROC SURVEYMEANS are produced using ODS (Output Delivery System). ODS encompasses more than just the production of output data sets. For a more detailed description on using ODS, see Chapter 15, "Using the Output Delivery System."

Displayed Output

By default PROC SURVEYMEANS displays a "Data Summary" table and a "Statistics" table. If you specify CLASS variables, or if you specify any character variables in the VAR statement, then the procedure displays a "Class Level Information" table. If you specify the LIST option in the STRATA statement, then the procedure displays a "Stratum Information" table.

Data and Sample Design Summary

The "Data Summary" table provides information on the input data set and the sample design. This table displays the total number of valid observations, where an observation is considered *valid* if it has nonmissing values for all procedure variables other than the analysis variables; that is, for all specified STRATA, CLUSTER, and WEIGHT variables. This number may differ from the number of nonmissing observations for an individual analysis variable, which the procedure displays in the "Statistics" table. See the section "Missing Values" on page 3197 for more information.

PROC SURVEYMEANS displays the following information in the "Data Summary" table:

- Number of Strata, if you specify a STRATA statement
- Number of Clusters, if you specify a CLUSTER statement
- Number of Observations, which is the total number of valid observations
- Sum of Weights, which is the sum over all valid observations, if you specify a WEIGHT statement

Class Level Information

If you use a CLASS statement to name classification variables for categorical analysis, or if you list any character variables in the VAR statement, then PROC SURVEYMEANS displays a "Class Level Information" table. This table contains the following information for each classification variable:

- Class Variable, which lists each CLASS variable name
- Levels, which is the number of values or levels of the classification variable
- Values, which lists the values of the classification variable. The values are separated by a white space character; therefore, to avoid confusion, you should not include a white space character within a classification variable value.

Stratum Information

If you specify the LIST option in the STRATA statement, PROC SURVEYMEANS displays a “Stratum Information” table. This table displays the number of valid observations in each stratum, as well as the number of nonmissing stratum observations for each analysis variable. The “Stratum Information” table provides the following for each stratum:

- Stratum Index, which is a sequential stratum identification number
- STRATA variable(s), which lists the levels of STRATA variables for the stratum
- Population Total, if you specify the TOTAL= option
- Sampling Rate, if you specify the TOTAL= option or the RATE= option. If you specify the TOTAL= option, the sampling rate is based on the number of valid observations in the stratum.
- N Obs, which is the number of valid observations
- Variable, which lists each analysis variable name
- N, which is the number of nonmissing observations for the analysis variable
- Clusters, which is the number of clusters, if you specify a CLUSTER statement,

Statistics

The “Statistics” table displays all of the statistics that you request with statistic-keywords described on page 3192, in the PROC SURVEYMEANS statement. If you do not specify any statistic-keywords, then by default this table displays the following information for each analysis variable: the sample size, the mean, the standard error of the mean, and the confidence limits for the mean. The “Statistics” table may contain the following information for each analysis variable, depending on which statistic-keywords you request:

- Variable name
- N, which is the number of nonmissing observations
- N Miss, which is the number of missing observations
- Minimum
- Maximum
- Range
- number of Clusters
- Sum of Weights
- DF, which is the degrees of freedom for the t test
- Mean
- Std Error of Mean, which is the standard error of the mean
- Var of Mean, which is the variance of the mean
- t Value, for testing H_0 : population MEAN $= 0$

- Pr $> \mid t \mid$, which is the two-sided p-value for the t test
- Lower and Upper $100(1-\alpha)$% CL for Mean, which are confidence limits for the mean
- Coeff of Variation, which is the coefficient of variation
- Sum
- Std Dev, which is the standard deviation of the sum
- Var of Sum, which is the variance of the sum
- Lower and Upper $100(1-\alpha)$% CL for Sum, which are confidence limits for the sum

ODS Table Names

PROC SURVEYMEANS assigns a name to each table it creates. You can use these names to reference the table when using the Output Delivery System (ODS) to select tables and create output data sets. These names are listed in the following table. For more information on ODS, see Chapter 15, “Using the Output Delivery System.”

Table 61.2. ODS Tables Produced in PROC SURVEYMEANS

ODS Table Name	**Description**	**Statement**	**Option**
ClassVarInfo	Class level information	CLASS	default
Statistics	Statistics	PROC	default
StrataInfo	Stratum information	STRATA	LIST
Summary	Data summary	PROC	default

By referring to the names of such tables, you can use the ODS OUTPUT statement to place one or more of these tables in output data sets.

For example, the following statements create an output data set named MyStrata, which contains the “StrataInfo” table, and an output data set named MyStat, which contains the “Statistics” table for the ice cream study discussed in the section “Stratified Sampling” on page 3186.

```
title1 ’Analysis of Ice Cream Spending’;
title2 ’Stratified Simple Random Sampling Design’;
proc surveymeans data=IceCream total=StudentTotal;
   stratum Grade / list;
   var Spending Group;
   ods output StrataInfo = MyStrata
              Statistics = MyStat;
run;
```

Examples

The “Getting Started” section on page 3183 contains examples of analyzing data from simple random sampling and stratified simple random sampling designs. This section provides more examples that illustrate how to use PROC SURVEYMEANS.

Example 61.1. Stratified Cluster Sample Design

Consider the example in the section “Stratified Sampling” on page 3186. The study population is a junior high school with a total of 4,000 students in grades 7, 8, and 9. Researchers want to know how much these students spend weekly for ice cream, on the average, and what percentage of students spend at least $10 weekly for ice cream.

The example in the section “Stratified Sampling” on page 3186 assumes that the sample of students was selected using a stratified simple random sampling design. This example shows analysis based on a more complex sample design.

Suppose that every student belongs to a study group and that study groups are formed within each grade level. Each study group contains between 2 and 4 students. Table 61.3 shows the total number of study groups for each grade.

Table 61.3. Study Groups and Students by Grade

Grade	Number of Study Groups	Number of Students
7	608	1,824
8	252	1,025
9	403	1,151
Total	617	4,000

It is quicker and more convenient to collect data from students in the same study group than to collect data from students individually. Therefore, this study uses a stratified clustered sample design. The primary sampling units, or clusters, are study groups. The list of all study groups in the school is stratified by grade level. From each grade level, a sample of study groups is randomly selected, and all students in each selected study group are interviewed. The sample consists of 8 study groups from the 7th grade, 3 groups from the 8th grade, and 5 groups from the 9th grade.

The SAS data set named IceCreamStudy saves the responses of the selected students.

```
data IceCreamStudy;
   input Grade StudyGroup Spending @@;
   if (Spending < 10) then Group=’less’;
     else Group=’more’;
   datalines;
7  34  7     7  34  7    7 412  4     9  27 14
7  34  2     9 230 15    9  27 15     7 501  2
9 230  8     9 230  7    7 501  3     8  59 20
7 403  4     7 403 11    8  59 13     8  59 17
8 143 12     8 143 16    8  59 18     9 235  9
8 143 10     9 312  8    9 235  6     9 235 11
9 312 10     7 321  6    8 156 19     8 156 14
7 321  3     7 321 12    7 489  2     7 489  9
7  78  1     7  78 10    7 489  2     7 156  1
7  78  6     7 412  6    7 156  2     9 301  8
;
```

In the data set IceCreamStudy, the variable Grade contain a student's grade. The variable StudyGroup identifies a student's study group. It is possible for students from different grades to have the same study group number because study groups are sequentially numbered within each grade. The variable Spending contains a student's response to how much he spends per week for ice cream, in dollars. The variable GROUP indicates whether a student spends at least $10 weekly for ice cream. It is not necessary to store the data in order of grade and study group.

The SAS data set StudyGroup is created to provide PROC SURVEYMEANS with the sample design information shown in Table 61.3.

```
data StudyGroups;
   input Grade _total_; datalines;
7 608
8 252
9 403
;
```

The variable Grade identifies the strata, and the variable _TOTAL_ contains the total number of study groups in each stratum. As discussed in the section "Specification of Population Totals and Sampling Rates" on page 3196, the population totals stored in the variable _TOTAL_ should be expressed in terms of the primary sampling units (PSUs), which are study groups in this example. Therefore, the variable _TOTAL_ contains the total number of study groups for each grade, rather than the total number of students.

The following SAS statements perform the analysis for this sample design.

```
title1 'Analysis of Ice Cream Spending';
title2 'Stratified Clustered Sample Design';
proc surveymeans data=IceCreamStudy total=StudyGroups;
   stratum Grade / list;
   cluster StudyGroup;
   var Spending Group;
run;
```

Output 61.1.1. Data Summary and Class Information

```
                      Analysis of Ice Cream Spending
                    Stratified Clustered Sample Design

                        The SURVEYMEANS Procedure

                               Data Summary

                   Number of Strata                   3
                   Number of Clusters                16
                   Number of Observations            40


                         Class Level Information

          Class
          Variable      Levels    Values

          Group              2    less more
```

Output 61.1.1 provides information on the sample design and the input data set. There are 3 strata in the sample design, and the sample contains 16 clusters and 40 observations. The variable Group has two levels, ’less’ and ’more’.

Output 61.1.2. Stratum Information

```
                      Analysis of Ice Cream Spending
                    Stratified Clustered Sample Design

                        The SURVEYMEANS Procedure

                           Stratum Information

Stratum               Population   Sampling
 Index       Grade         Total       Rate      N Obs   Variable              N
--------------------------------------------------------------------------------
   1             7           608       0.01         20   Spending             20
                                                         Group = less         17
                                                         Group = more          3
   2             8           252       0.01          9   Spending              9
                                                         Group = less          0
                                                         Group = more          9
   3             9           403       0.01         11   Spending             11
                                                         Group = less          6
                                                         Group = more          5
--------------------------------------------------------------------------------

                           Stratum Information

Stratum               Population   Sampling
 Index       Grade         Total       Rate      N Obs   Variable       Clusters
--------------------------------------------------------------------------------
   1             7           608       0.01         20   Spending              8
                                                         Group = less          8
                                                         Group = more          3
   2             8           252       0.01          9   Spending              3
                                                         Group = less          0
                                                         Group = more          3
   3             9           403       0.01         11   Spending              5
                                                         Group = less          4
                                                         Group = more          4
--------------------------------------------------------------------------------
```

Output 61.1.2 displays information for each stratum. Since the primary sampling units in this design are study groups, the population totals shown in Output 61.1.2 are the total numbers of study groups for each stratum or grade. This differs from Figure 61.3 on page 3187, which provides the population totals in terms of students since students were the primary sampling units for that design. Output 61.1.2 also displays the number of clusters for each stratum and analysis variable.

Output 61.1.3. Statistics

```
                           Analysis of Ice Cream Spending
                         Stratified Clustered Sample Design

                             The SURVEYMEANS Procedure

                                     Statistics

                                                Std Error     Lower 95%     Upper 95%
Variable                  N          Mean         of Mean   CL for Mean   CL for Mean
-------------------------------------------------------------------------------------
Spending                 40      8.750000        0.634549      7.379140     10.120860
Group = less             23      0.575000        0.056274      0.453427      0.696573
Group = more             17      0.425000        0.056274      0.303427      0.546573
-------------------------------------------------------------------------------------
```

Output 61.1.3 displays the estimates of the average weekly ice cream expense and the percentage of students spending at least $10 weekly for ice cream. These estimates are the same as those shown in Figure 61.4 for the stratified SRS design. However, the variance estimates are different because of the different sample designs.

Example 61.2. Unequal Weighting

Quite often in complex surveys, respondents have unequal weights, which reflect unequal selection probabilities and adjustments for nonresponse and poststratification. In such surveys, the appropriate sampling weights must be used to obtain valid estimates for the study population. This example illustrates analysis of survey data with unequal sampling weights.

Suppose that economists want to study profiles of the 800 top-performing companies to provide information on their impact on the economy. A sample of 66 companies is selected with unequal probability.

```
data Company;
   length Type $14;
   input Type$ Asset Sale Value Profit Employee Weight;
   datalines;
Other            2764.0  1828.0  1850.3   144.0   18.7   9.6
Energy          13246.2  4633.5  4387.7   462.9   24.3  42.6
Finance          3597.7   377.8    93.0    14.0    1.1  12.2
Transportation   6646.1  6414.2  2377.5   348.2   47.1  21.8
HiTech           1068.4  1689.8  1430.2    72.9    4.6   4.3
Manufacturing    1125.0  1719.4  1057.5    98.1   20.4   4.5
Other            1459.0  1241.4   452.7    24.5   20.1   5.5
Finance          2672.3   262.5   296.2    23.1    2.2   9.3
Finance           311.0   566.2   932.0    52.8    2.7   1.9
```

```
Energy           1148.6  1014.6   485.1    60.6    4.0   4.5
Finance          5327.0   572.4   372.9    25.2    4.2  17.7
Energy           1602.7   678.4   653.0    75.6    2.8   6.0
Energy           5808.8  1288.4  2007.0   318.8    5.9  19.2
Medical           268.8   204.4   820.9    45.6    3.7   1.8
Transportation   5222.6  2627.8  1910.0   245.6   22.8  17.4
Other             872.7  1419.4   939.3    69.7   12.2   3.7
Retail           4461.7  8946.8  4662.7   289.0  132.1  15.0
HiTech           6719.2  6942.0  8240.2   381.3   85.8  22.1
Retail            833.4  1538.8  1090.3    64.9   15.4   3.5
Finance           415.9   167.3  1126.8    56.8    0.7   2.2
HiTech            442.4  1139.9  1039.9    57.6   22.7   2.3
Other             801.5  1157.0   664.2    56.9   15.5   3.4
Finance          4954.8   468.8   366.4    41.7    3.0  16.5
Finance          2661.9   257.9   181.1    21.2    2.1   9.3
Finance          5345.8   530.1   337.4    36.4    4.3  17.8
Energy           3334.3  1644.7  1407.8   157.6    6.4  11.4
Manufacturing    1826.6  2671.7   483.2    71.3   25.3   6.7
Retail            618.8  2354.7   767.7    58.6   19.0   2.9
Retail           1529.1  6534.0   826.3    58.3   65.8   5.7
Manufacturing    4458.4  4824.5  3132.1    28.9   67.0  15.0
HiTech           5831.7  6611.1  9464.7   459.6   86.7  19.3
Medical          6468.3  4199.2  3170.4   270.1   59.5  21.3
Energy           1720.7   473.1   811.1    86.6    1.6   6.3
Energy           1679.7  1379.9   721.1    91.8    4.5   6.2
Retail           4018.2 16823.4  2038.3   178.1  162.0  13.6
Other             227.1   575.8  1083.8    62.6    1.9   1.6
Finance          3872.8   362.0   209.3    27.6    2.4  13.1
Retail           3359.3  4844.7  2651.4   224.1   75.6  11.5
Energy           1295.6   356.9   180.8   162.3    0.6   5.0
Energy           1658.0   626.6   688.0   126.0    3.5   6.1
Finance         12156.7  1345.5   680.7   106.6    9.4  39.2
HiTech           3982.6  4196.0  3946.8   313.9   64.3  13.5
Finance          8760.7   886.4  1006.9    90.0    7.5  28.5
Manufacturing    2362.2  3153.3  1080.0   137.0   25.2   8.4
Transportation   2499.9  3419.0   992.6    47.2   25.3   8.8
Energy           1430.4  1610.0   664.3    77.7    3.5   5.4
Energy          13666.5 15465.4  2736.7   411.4   26.6  43.9
Manufacturing    4069.3  4174.7  2907.6   289.2   38.2  13.7
Energy           2924.7   711.9  1067.8   146.7    3.4  10.1
Transportation   1262.1  1716.0   364.3    71.2   14.5   4.9
Medical           684.4   672.9   287.4    61.8    6.0   3.1
Energy           3069.3  1719.0  1439.0   196.4    4.9  10.6
Medical           246.5   318.8   924.1    43.8    3.1   1.7
Finance         11562.2  1128.5   580.4    64.2    6.7  37.3
Finance          9316.0  1059.4   816.5    95.9    8.0  30.2
Retail           1094.3  3848.0   563.3    29.4   44.7   4.4
Retail           1102.1  4878.3   932.4    65.2   47.3   4.4
HiTech            466.4   675.8   845.7    64.5    5.2   2.4
Manufacturing   10839.4  5468.7  1895.4   232.8   47.8  35.0
Manufacturing     733.5  2135.3    96.6    10.9    2.7   3.2
Manufacturing   10354.2 14477.4  5607.2   321.9  188.5  33.5
Energy           1902.1  2697.9   329.3    34.2    2.2   6.9
Other            2245.2  2132.2  2230.4   198.9    8.0   8.0
```

```
Transportation    949.4  1248.3   298.9    35.4   10.4   3.9
Retail           2834.4  2884.6   458.2    41.2   49.8   9.8
Retail           2621.1  6173.8  1992.7   183.7  115.1   9.2
;
```

The variable Type identifies the type of market for the company. The variable Asset contains the company's assets in millions of dollars. The variable Sale contains sales in millions of dollars. The variable Value contains the market value of the company in millions of dollars. The variable Profit contains the profit in millions of dollars. The variable Employee stores the number of employees in thousands, and the variable Weight contains the sampling weight. In this example, the sampling weights are reciprocals of the selection probabilities.

Using a probability sample design and the appropriate sampling weights, you can obtain statistically valid estimates for the study population. The following SAS statements compute estimates for this study.

```
title1 'Top Companies Profile Study';
title2 'Using Sampling Weights';
proc surveymeans data=Company total=800 mean sum;
   var Asset Sale Value Profit Employee;
   weight Weight;
run;
```

The TOTAL=800 option specifies the total number of companies in the study population. The statistic-keywords MEAN and SUM request estimates of the mean and total for the analysis variables. The WEIGHT statement identifies the sampling weight variable Weight. The VAR statement lists the variables to analyze.

Output 61.2.1. Company Profile Study

```
                          Top Companies Profile Study
                             Using Sampling Weights

                           The SURVEYMEANS Procedure

                                  Data Summary

                  Number of Observations           66
                  Sum of Weights                799.8


                                   Statistics

                                   Std Error
Variable            Mean           of Mean               Sum         Std Dev
------------------------------------------------------------------------
Asset        6523.488510        720.557075           5217486         1073829
Sale         4215.995799        839.132506           3371953          847885
Value        2145.935121        342.531720           1716319          359609
Profit        188.788210         25.057876            150993           30144
Employee       36.874869          7.787857             29493     7148.003298
------------------------------------------------------------------------
```

Output 61.2.1 shows that there are 66 observations in the sample. The sum of the sampling weights equals 799.8, which is close to the total number of companies in the study population.

The “Statistics” table in Output 61.2.1 displays the estimates of the mean and total for all analysis variables.

If you do not use the appropriate sampling weights, then the results of the analysis may be biased. For example, the following statements analyze the data without the sampling weights that reflect the unequal probabilities of selection.

```
title1 ’Top Companies Profile Study’;
title2 ’Without Using the Sampling Weights’;
proc surveymeans data=Company total=800 mean sum;
   var Asset Sale Value Profit Employee;
run;
```

Output 61.2.2. Company Profile Study without Sampling Weights

```
                         Top Companies Profile Study
                      Without Using the Sampling Weights

                          The SURVEYMEANS Procedure

                                 Data Summary

                     Number of Observations            66


                                  Statistics

                                  Std Error
Variable            Mean            of Mean             Sum          Std Dev
------------------------------------------------------------------------
Asset        3557.753030         401.508963          234812            26500
Sale         2881.306061         407.864339          190166            26919
Value        1517.507576         206.197430          100156            13609
Profit        129.121212          13.824279     8522.000000       912.402420
Employee       27.704545           4.601392     1828.500000       303.691848
------------------------------------------------------------------------
```

Output 61.2.2 shows the results of the analysis without using the sampling weights. These results just summarize the sample of 66 companies, and they are not statistically valid estimates for the study population of 800 companies. These statistics are substantially different from the estimates shown in Output 61.2.1. For example, the total assets computed without sampling weights is only $235 billion, compared to the estimate of $5,217 billion computed with the sampling weights. The estimated mean of the assets is $3.56 billion for the sample, but is estimated as $6.52 billion for the study population.

References

Cochran, W. G. (1977), *Sampling Techniques*, Third Edition, New York: John Wiley & Sons, Inc.

Foreman, E. K. (1991), *Survey Sampling Principles*, New York: Marcel Dekker, Inc.

Fuller, W. A. (1975), "Regression Analysis for Sample Survey," *Sankhyā*, 37, Series C, Pt. 3, 117–132.

Hansen, M. H., Hurwitz, W. N., and Madow, W. G. (1953), *Sample Survey Methods and Theory*, Volumes I and II, New York: John Wiley & Sons, Inc.

Kalton, G. (1983), *Introduction to Survey Sampling*, SAGE University Paper series on Quantitative Applications in the Social Sciences, series no. 07-035, Beverly Hills and London: SAGE Publications, Inc.

Kish, L. (1965), *Survey Sampling*, New York: John Wiley & Sons, Inc.

Lee, E. S., Forthoffer, R. N., and Lorimor, R. J. (1989), *Analyzing Complex Survey Data*, Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-071, Beverly Hills and London: Sage Publications, Inc.

Pringle, R. M. and Raynor, A. A. (1971), *Generalized Inverse Matrices with Applications to Statistics*, New York: Hafner Publishing Co.

Statistical Laboratory (1989), *PC CARP*, Ames, IA: Statistical Laboratory, Iowa State University.

Woodruff, R. S. (1971), "A Simple Method for Approximating the Variance of a Complicated Estimate," *Journal of the American Statistical Association*, 66, 411–414.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS/STAT® User's Guide, Version 8*, Cary, NC: SAS Institute Inc., 1999.

SAS/STAT® User's Guide, Version 8

ISBN 1–58025–494–2

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st printing, October 1999

The Institute is a private company devoted to the support and further development of its software and related services.